AD_____

Award Number:
W81XWH-08-1-0402

TITLE:
Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer

PRINCIPAL INVESTIGATOR:
William D Foulkes MB PhD

CONTRACTING ORGANIZATION:
Jewish General Hospital
Montreal, Quebec
H3T 1E2

REPORT DATE:
July 2010

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-07-2010	Annual	01-07-2009 to 30-06-2010

4. TITLE AND SUBTITLE

Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER

W81XWH-08-1-0402

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Foulkes, William D.
Email: william.foulkes@mcgill.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Jewish General Hospital
Montreal, QC

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command, Fort Detrick, MD, 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this proposal, we set out to a) systematically monitor splicing variant profiles in breast cancer susceptibility genes and b) explore the role of alternative splicing in breast chemotherapy using a global strategy. In doing so, we hope to identify and validate candidate splicing variants involved in tumorigenesis using polony digital exon-profiling and functional assays. We are moving forward on four fronts – 1) the barcode methodology is in development; 2) we are working with state-of-the art capture arrays; 3) we are using the very latest RNA sequencing technology and 4) we are conducting comprehensive analyses of existing splice site variants in known breast cancer susceptibility genes. In part 3, we have generated over 5 million reads that have been aligned to the splice junction libraries, and we are using these reads to quantify and characterize alternative splicing events. Moreover, we will add value to this project by attempting to identify fusion proteins. In part 4, we have determined that the BRCA2 isoform known as BRCA2Δex12 is not associated with a recognizable phenotype. This work is now in press in Human Mutation.

15. SUBJECT TERMS

Breast cancer, splicing, single molecule analysis, RNA sequencing, BRCA2, Variants of Unknown Significance, Functional assays

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	15	19b. TELEPHONE NUMBER *(Include area code)*

Table of Contents

Page

Exploring the Pathogenic and Therapeutic Implications of Aberrant Splicing in Breast Cancer

W81XWH-08-1-0402

PRINCIPAL INVESTIGATORS: William D Foulkes MB PhD, Jun Zhu PhD

CO-INVESTIGATOR: Jacek Majewski, PhD

Introduction

In this proposal, we aimed to a) systematically monitor splicing variant profiles in breast cancer susceptibility genes and b) explore the role of alternative splicing in breast chemotherapy using a global strategy. In doing so, we hoped to identify and validate candidate splicing variants involved in tumorigenesis using polony digital exon-profiling and functional assays. This annual report summarizes our progress in the second year of this two-year award.

Body

Verbatim – from S.O.W

Objective 1: To profile splicing patterns of 100 breast cancer-related genes in 30 normal and tumor

breast cell lines

Objective 2: To profile splicing patterns of 100 cancer-related genes in lymphocytes and breast tissues

Objective 3: Data analysis and validation to determine the role of aberrant splicing in breast tumourgenesis *(months 8-12)*

These objectives can be summarized in terms of the expected Results/Deliverables from Year 2: 1) Optimizing the method of bar-code PCR-sequencing for studying alternative splicing at candidate loci. 2) Identifying splicing variants associated with increased breast cancer susceptibility derived from the candidate loci. Year 2 focuses more on drug sensitivity, and how splicing patterns may determine the response to certain chemotherapeutic agents.

We have been attempting to optimize the method of bar-code PCR-sequencing for studying alternative splicing at candidate loci. This part of the project is the responsibility of Dr. Zhu, and is covered in detail in his report. As stated in the first annual report, we switched to an emerging alternative strategy, NimbleGen Sequence-Capture array that can selectively enrich genes of interest (3-20 Mbps) for deep sequencing analysis. The captured fragments were then subjected to 454 pyrosequencing to obtain a longer read length (~400bp), which enables better identification of alternative splicing and mutation detection.

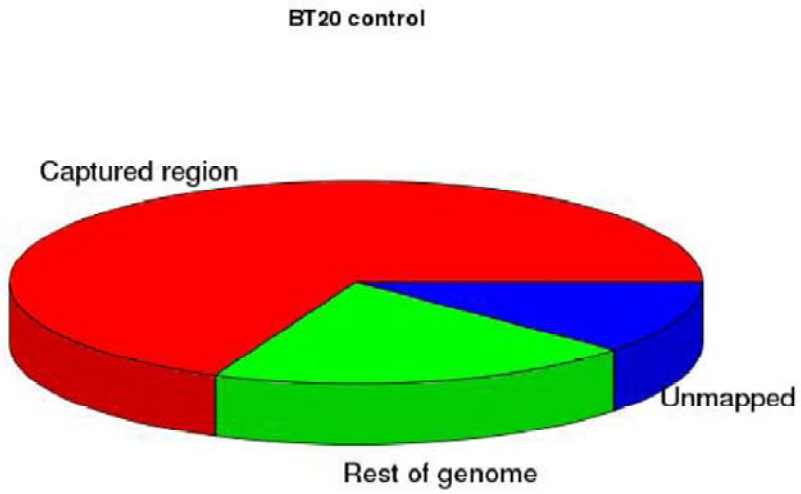
This work has now been completed by Ting Ni (post-doctoral fellow, funded by this award, based at Duke University). We constructed two libraries with BT20 cell line and its derivative that resistant to

dastanib, an important new tyrosine kinase drug that has considerable activity in leukemia and in some solid tumors, including breast cancer. Two libraries for Illumina platform and two libraries for 454 sequencing have been created. The progress on this front can be found in Dr. Jun Zhu's annual progress report.

In addition to the analysis of the BT20 lines, through collaboration with the Breakthrough Breast Cancer Institute in the UK, we have collected RNA samples from 34 breast tumor lines that represent 5 different types of breast cancers classified by expression profiles: luminal epithelial-like subtypes A and B, basal epithelial-like, ERBB2-amplification associated and normal breast-like subtypes. The fact that the response profiles of these lines to dasatinib have been published, will allow us to assess whether or not the genetic alterations identified in an induced resistance model can be applied to intrinsic resistance. Thus, as soon as our pipeline is well-developed we will be able to test and validate candidate genes using this well-characterised set of breast cancer cell lines.

First of all, we showed that most of the reads mapped to the captured region in a) the control line (Figure 1) and in b) the line with induced resistance to dasatinib (Figure 2).

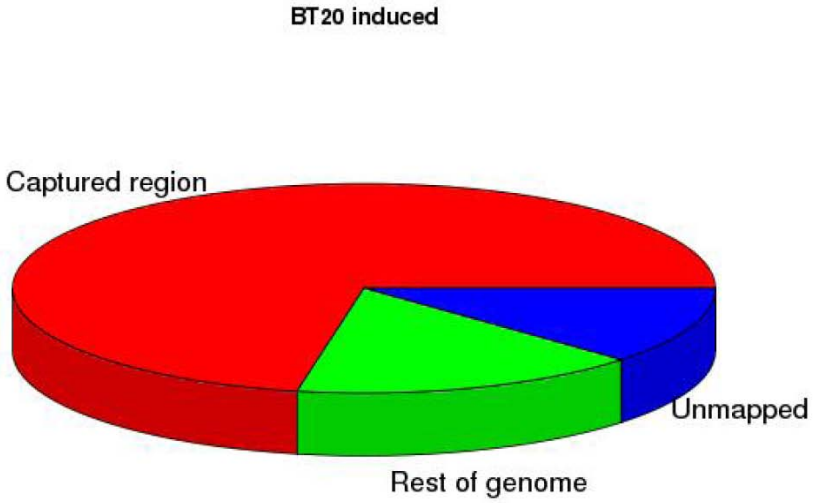
Figure 1



Sequences mapping to the captured region are now being analyzed and compared with dbSNP and other reference databases.

We have identified many promising splice site mutations, indels and SNPs that we are currently investigating. Notably, our collaborator Dr. Raquel Aloyz has been working on the cell biology of the differences between these two lines, and we hope to integrate the datasets soon.

Figure 2



The third strategy we have used to characterize alternative splicing events is to sequence the RNA of breast cancer cell lines and one breast tumor directly. Because the pace of technological innovation is substantially faster than the granting process, our original proposal to analyze the splicing patterns of 100 candidate breast cancer genes (using the barcode and/or capture array strategy) has been augmented by state-of-the art whole transcriptome sequencing or RNA-Seq. In other words, in this part of the project, we aim to analyze not the 100 "top" candidate breast cancer susceptibility genes, but instead all ~24,000 genes. This project has been led by PI Foulkes and co-PI Majewski. To date, we have used Illumina/Solexa RNA sequencing to generate millions of raw reads from the tumor and control lines indicated in Table 1 below. These reads were mapped to the transcriptome using various alignment tools including BWA, TopHat and ELAND and were visualized using the Integrated Genome Viewer from the Broad Institute. In the single-end sequencing strategy, a single read is generated per transcriptome fragment. On the other hand, in paired-end strategy, the fragment is sequenced from both ends, creating a pair of reads with an expected distance apart.
We have chosen RNA-Seq for the following reasons –

- It combines the accuracy and sensitivity of Sanger sequencing with the throughput of microarrays
    - It characterizes the transcriptome at a single-base resolution, producing tens of millions of short read sequences per sample
- It measures gene and isoform expression levels and provides structural information
- It is able to reveal:
    - SNPs and indels
    - Allelic expression
    - Gene fusion and novel gene isoform transcripts
    - Gene expression

We have chosen to focus on the *BRCA1* transcriptome because we have access to quality samples and because we felt that in a highly competitive field, we needed to find a niche. At the current time, whole transcriptome sequencing is too expensive to allow us to stick to our original plan, so instead of

studying 100 genes in many samples, we have opted to study all genes in few samples. As prices fall, we will be able to study more samples.

Here, we intend to -

- Compare *BRCA1*-deficient breast cancer samples to controls
- Characterize the *BRCA1* breast cancer transcriptome via next-generation RNA sequencing (RNA-seq) by identifying:
    - fusion proteins – a type of spliced protein where exons from two different genes are spliced together to make a new protein, with potentially new functions.
    - new splice isoforms - the main objective of the current proposal
      SNPs and indels - this is not the main thrust of the current proposal, but since we have these data, we will be able to study these at a later date

Positive results can provide new biomarkers and detect new proteins or pathways involved in *BRCA1*-related breast cancer formation

Table 1: Cell lines and tumors used in this study. The samples with purple background are our *BRCA1*-deficient samples whereas those with blue background are our control samples.

Sample ID	Description	BRCA1 mutation	Read length
HCC1937‡	Breast cancer cell line	5266dupC	50bp single-end
SUM149PT*	Breast cancer cell line	2288delT	36bp paired-end
SUM131502‡	Breast cancer cell line	68_69delAG	36bp paired-end
HCC3153‡	Breast cancer cell line	943ins10	36bp paired-end
T92‡	Breast cancer tumor	5266dupC	76bp paired-end
HCC2337‡	Matched lymphocyte from same patient as HCC1937	5266dupC	76bp paired-end
MCF10A‡	Non-tumourigenic breast epithelial cell line	N/A	76bp paired-end

Sequenced at ‡McGill University and Genome Quebec Innovation Centre, Canada,
*Institute of Cancer Research, UK

The millions of RNA-Seq reads are aligned to splice junction libraries, and we use these reads to quantify and characterize alternative splicing events. However, one of our main interests was to identify novel splice isoforms (not present in most reference splice libraries). To this end, we have pursued

three approaches: 1) Mapping sequencing reads to exhaustive splice junction libraries representing all possible splice junctions *within* each gene; 2) creating heuristic algorithms for mapping reads to an exhaustive library joining all exons *across* all genes in the human genome 3) monitoring excess reads that map to normally intronic sequences – this is meant to detect novel intron retention events;. The second approach will allow us to detect all known and novel alternative splicing events, trans-splicing (across genes), as well gene fusion events which are known to occur in many tumors; . In other words, we are not just looking for known splicing events, but all possible combinations of exons within a given gene as well as between any two genes in the genome.

A) Splicing results

First, we present our results from the splicing analysis of the cell lines/tumor in Table 1 above. Note the very large number (>1400) of novel junctions identified in at least two of the breast cancer samples yet absent or with very low expression in the control samples (Figures 3 and 4). We ignore events which were found in a single breast cancer sample as our aim is to identify novel splice variants which may characterize *BRCA1* breast cancers and as such must be recurrent. In Figure 4, we distinguish between coding and noncoding (ie, located within 3' or 5' UTRs) since these categories of events have different impacts on the gene in question. Coding events can introduce a frameshift and possibly a premature termination codon (PTC) whereas noncoding events can affect the stability of the resulting mRNA and/or introduce a novel transcription start site and potentially a frameshift. Both types of events have the potential to generate a truncated protein and possibly induce nonsense-mediated decay, depending on the location of the alternative splicing event within the gene in question.

Figure 3: Types of splice variants that we have observed in the tumors/cell lines studied in Table 1. ***NB*** Alternative 3'SS or 5'SS in known exons refers to events in which the splice site is shifted by 1-9bp whereas any splice site that is shifted by more than 9bp is considered a novel exon. Unannotated regions refer to nongenic regions in the NCBI hg18 reference.

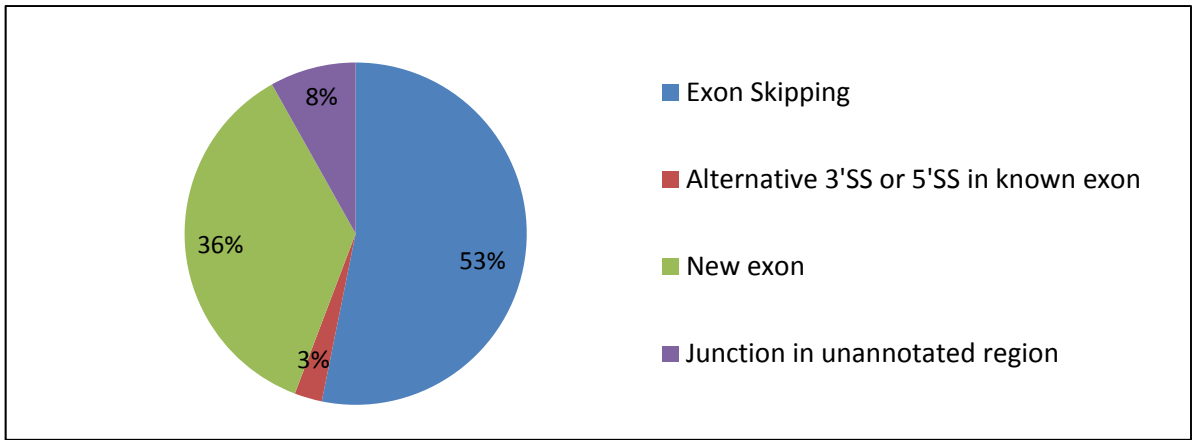
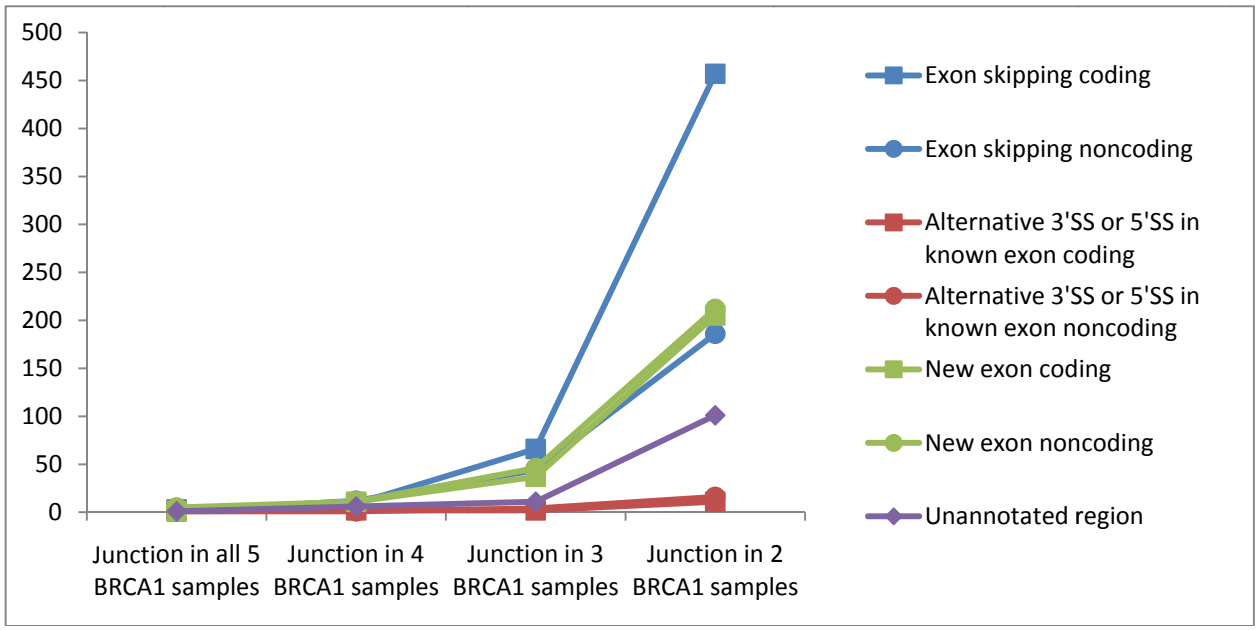


Figure 4: Breakdown of the types of splice variants seen per number of breast cancer samples.
NB These junctions had very low expression, if any, in the HCC2337 and MCF10A control cell lines. Noncoding exons refer to exons located in the 3' or 5' UTRs.



Moreover, a number of these novel splice junctions are biologically intriguing, no more so than the novel isoform of *PALB2* that we have identified (Figure 5). This requires validation, but since the new isoform results in a truncated protein and truncating *PALB2* mutations are associated with breast cancer, the finding is of considerable interest. The exon skipping event leads to the predicted obliteration of a WD40 domain in *PALB2*, an essential domain for binding of BRCA2. The PALB2 and BRCA2 protein interaction is important for cell cycle control and apoptosis.

Figure 5: Example of an interesting novel junction identified by our approach.

Figure 5A shown below is a screen shot from the Integrated Genome Viewer (Broad Institute) illustrating the alignment within the end of the *PALB2* gene. A significant number of reads are shown which skip over exon 12. This event leads to a frameshift and a premature termination codon.

Figure 5A

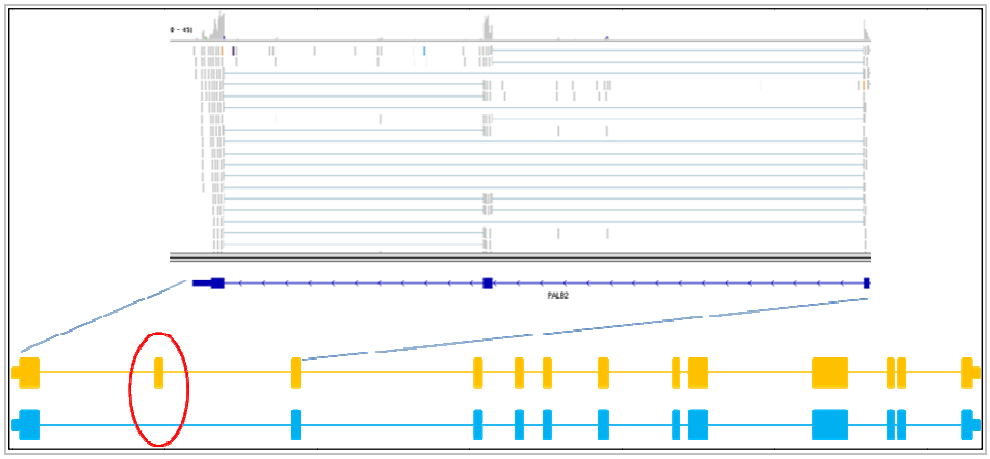


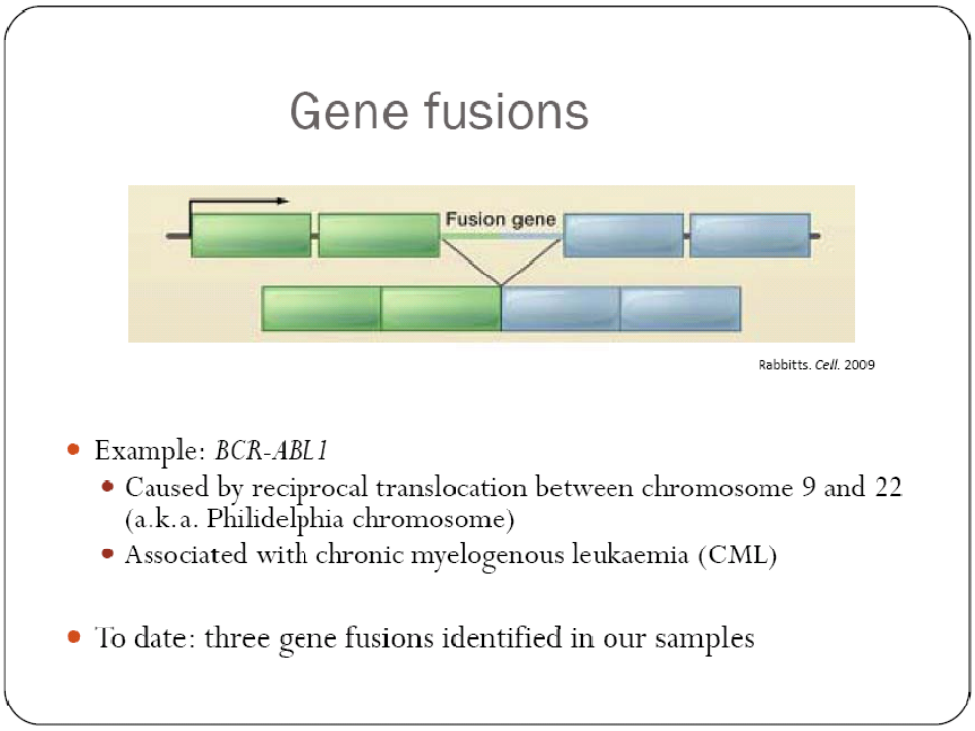
Figure 5B

Figure 5B shows that the truncated isoform loses the portion coding for one of the WD40 domains (see text).

B) Fusion protein project – rationale and results

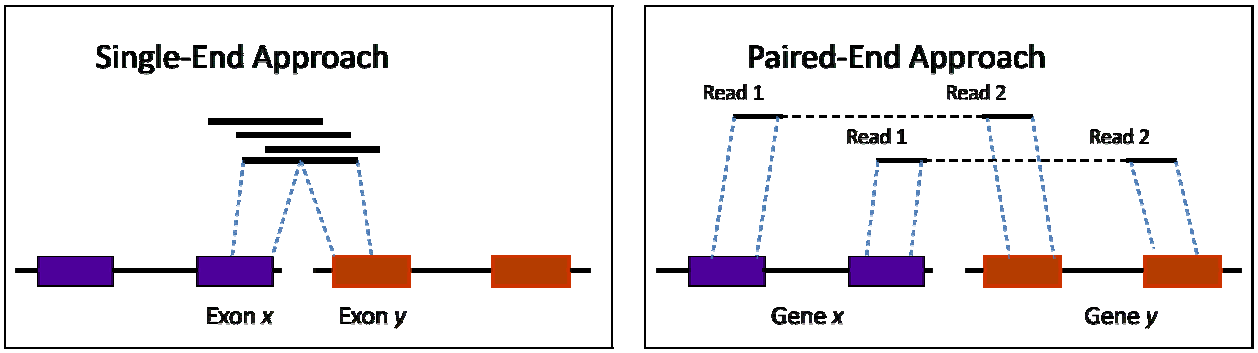
Gene fusions are the result of chromosomal alterations involving two genes. These chimeras may have severe phenotypic effects, such as the well-studied *BCR-ABL1* fusion protein implicated in chronic myelogenous leukemia (Shtivelman et al., 1985) and *TMPRSS2-ERG* found in many cases of prostate cancer (Tomlins et al., 2005). New efforts using high-throughput sequencing have resulted in new discoveries of gene fusions. This has prompted interest in determining whether these chromosomal aberrations may be specific to cancer and if they are, may theoretically serve as an ideal diagnostic and therapeutic target (Prensner and Chinnaiyan, 2009).

Figure 6: Gene fusions - examples and summary of our results



We have devised two strategies to identify gene fusions using either single-end or paired-end data, following a similar approach described by Maher et al. (2009) (Figure 7). In both cases, we downloaded from the UCSC Genome Database (hg18 assembly) the set of exon genomic sequences from all mature mRNA RefSeq transcripts. As the mRNA of gene fusions would typically involve the fusion of exon sequences from two different genes, we retained only the boundary sequences of each exon (i.e. 49 bp sequence from the left boundary of an exon and 49 bp sequence from the right boundary of an exon, for 50bp reads), in order to identify reads that may span such exon-exon boundaries.

Figure 7. Discovering gene fusions using single-end and paired-end RNA-Seq datasets.



The basic principle for single-end data is described below using the HCC1937 dataset as an example. Of a total of 40 million 50 bp reads sequenced from the this cell line, approximately 1.5 million reads were

determined to be non-mapping after analyzing the data with BWA. Non-mapping reads are those which did not have a unique alignment to the reference genome. We limited our single-end gene fusion analysis to only these reads as they may potentially characterize breakpoints of gene fusions not found in the reference genome. Blat (Kent, 2002) was used to align the non-mapping reads against the list of exon boundary sequences. Using a set of in-house scripts written in Python, we filtered the results for alignments such that a read was partially aligning to an exon boundary. Of these partial alignments, we further identified whether its remaining unaligned sequence aligned to another exon region either from the Blat analysis or by simple string matching techniques. This methodology will identify gene fusions with breakpoints located within introns but not necessarily those with breakpoints in exons. However, since exons make up only 5% of the genome, we assume that the breakpoint events occur within the intron. Moreover, if a breakpoint does occur within an exon, the coding frame would have to be maintained in order for the fusion to be expressed.

For paired-end reads, we took advantage of the additional distance information. As previously mentioned, each paired reads should map at an expected distance from each other in the transcriptome. Thus, potential gene fusion events can be inferred by paired reads which map on two different mRNA corresponding to two different genes (see Figures 8 and 9). The genes may be located on the same chromosome (implying an intrachromosomal rearrangement) or different chromosome (interchromosomal). Thus, candidate gene fusion events are nominated based on satisfying two criteria. First, we look for a sufficient amount of supporting paired reads that map unusually in two specific loci. Secondly, we then generate potential exon-exon junction sequences joining the two genes and search for additional individual reads that map across the two exons. The latter step is analogous to the previously described single-end approach and provides further supporting evidence for the candidate gene fusion.

We first implemented the single-end procedure and tested it on HCC1937, where we successfully identified the *NFIA*-*EHF* gene fusion. This fusion was previously identified by whole-genome DNA sequencing and validated by RT-PCR and FISH in a study by Stephens et al. (*Nature*, 2009). Hence, we demonstrated as proof of principle that we can independently identify the same gene fusion using RNA-Seq. We have now studied all the lines/tumors in table 1 and identified two additional novel gene fusions (Figures 8 and 9).

B) Gene fusions detected

Figure 8: Novel fusion protein identified in SUM149.

NB The top five reads supporting the fusion are evidence resulting from the paired-end strategy whereas the bottom three reads are results of the single-end strategy.

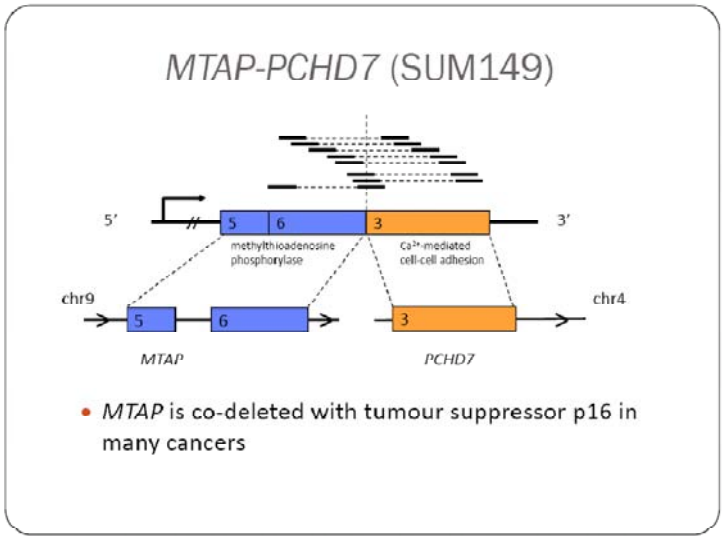
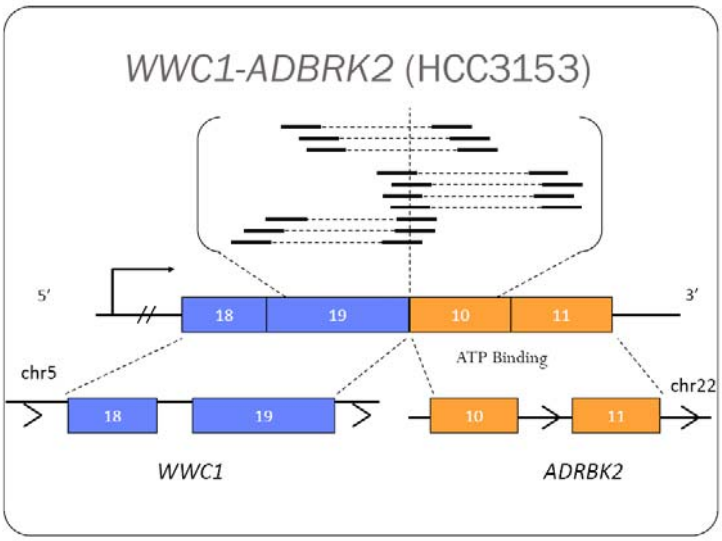


Figure 9: Novel fusion protein identified in HCC3153.

NB The top three reads supporting the fusion are evidence resulting from the paired-end strategy whereas the bottom seven reads are results of the single-end strategy.



NB this fusion protein has been confirmed by RT-PCR and Sanger sequencing of RNA.

Summary of progress

Thus, although the original the molecular barcode strategy has not yet been implemented (see report of co-PI, Dr Zhu), we are have made significant findings in three areas –1) we are working with state-of-the art capture arrays in the sample BT-20 (with and without induced resistance to dasatininb) ; 2) we have identified over 1400 new splice variants and 3) We have identified novel fusion proteins.

Key Research Accomplishments

- Publication from the Foulkes group _ Human Mutation, 2009
- Development of novel approaches to analyzing RNA Seq data
- Completion of BT-20 Nimblegen array
- Identification of >1400 novel splice site junctions in breast cancers and breast cancer cell lines
- Identification of 2 novel fusion proteins in breast cancer cell lines

Reportable Outcomes

- Abstract presented at Cold Spring Harbor Meeting (Kevin Ha, first author)
- Abstract submitted to American Society of Human Genetics Meeting (Emilie Lalonde, first authors)
- Data presented at McGill University Graduate Research Day (joint first prize awarded to Emilie Lalonde

NB Kevin Ha is the graduate student of Dr. Majewski and Emilie Lalonde is the graduate student of Dr. Majewski and Foulkes (co-supervision).

Conclusion

In this proposal, we have set out to evaluate the importance of splicing for breast cancer biology. We are employing a number of state-of-the-art technologies – a custom-made capture array to study splicing events in breast cancer and RNA sequencing, which is a relatively unbiased approach to the problem. We hope that our approach will lead to significant insights into the more general question of the importance of alternative splicing in breast cancer biology.

References

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res, 12(4), 656-664.

Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., et al. (2009). Transcriptome sequencing to detect gene fusions in cancer. Nature, 458(7234), 97-101.

Prensner, J. R., & Chinnaiyan, A. M. (2009). Oncogenic gene fusions in epithelial carcinomas. Curr Opin Genet Dev, 19(1), 82-91.

Shtivelman, E., Lifshitz, B., Gale, R. P., & Canaani, E. (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature, 315(6020), 550-554.

Stephens PJ, McBride DJ, Lin ML, et a. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009 Dec 24;462(7276):1005-10.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X., et al. (2005). Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science, 310(5748), 644-648.